DOCUMENT RESUME

ED 286 914 TM 870 546

AUTHOR Sanber, Shukri R.; Millman, Jason
TITLE Gender and Race Effects on Standardized Tests Predictive Validity: A Meta-analytical Study.
PUB DATE Apr 87
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143) -- Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; Black Students; Females; *Grade Prediction; Higher Education; Males; Meta Analysis; *Predictive Validity; *Racial Differences; Regression (Statistics); Research Problems; *Sex Differences; *Test Validity; White Students

ABSTRACT

This paper summarizes and analyzes data from studies of differential validity and differential prediction for blacks versus whites and females versus males. Forty studies on racial differences and 53 on gender differences were selected from searches of the ERIC, PsychInfo, and Dissertation Abstracts International databases; Buros' Mental Measurements Yearbooks; and major test reviews. All data were derived from standardized achievement tests. Two questions were investigated: (1) is there empirical evidence of differential predictive validity on race and/or gender?; and (2) is there an under- or over-prediction of black or female students' performance when combined regression equations are used? Results of the meta-analysis showed: (1) averaged validity coefficients were higher for whites than blacks, and for females than males, but the differences were too small to be meaningful; (2) averaged regression coefficients were higher for whites than blacks and for males than females; and (3) slopes were equal in 78% of the black/white comparisons and 81% of the male/female comparisons. In 41% of these black/white comparisons, intercepts were higher for blacks, and in 53% of the male/female comparisons intercepts were higher for females. It was concluded that, despite small differences in regression equations between the racial and gender groups, tests as predictors of future academic performance areas valid for blacks and females as they are for whites and males. (A list of studies included in the synthesis and a six-page bibliography are included.) (JGL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED286914

Gender and Race Effects on Standardized Tests Predictive Validity: A Meta-analytical Study.

Shukri R. Sanber, Bethlehem University
and
Jason Millman, Cornell University.

1987 AERA Annual Meeting
Washington, D.C.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. Sanber

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM 870 546

BEST COPY AVAILABLE

Students' scores on standardized tests are used either by themselves or in conjunction with other indicators for decision making in diversified educational settings such as tracking, placement, and admissions. Since these decisions are usually of great social importance, tests have attracted the attention of scholars, practioners, and politicians since their initiation in the early 1900's. One major and consistent challenge to the use of standardized tests in schools and colleges has been the claim of test bias in the form of differential predictive validity relative to the different subgroups that compose the population of students. Bias in testing can be examined at different levels. It can be studied at the item level. A good treatment and review of item bias is Berk's (1982).

Bias in testing can be viewed as a question of fairness, which is a philosophical concept that has more to do with social policies (Cole, 1981; Schunneman, 1981). For a good treatment on this issue see Van der Kamp (1980). Neither of these views is treated in this study. The concern here is the differential level of relationship between test scores and later measures of performance for subgroups of students.

Scores of empirical studies have been conducted to investigate the relevance of standardized tests for all subgroups. Each of these studies has been typically constrained by spatial and temporal limitations, due to logistic reasons. Therefore a literature review of many studies has become the only logical approach to draw broad conclusions. Most of the well-known contributions in the area of differentail predictive

validity in educational settings (e.g. Linn, 1982; Scheunneman, 1981; and Jensen, 1980) have been based upon narrative literature reviews.

However, due to the selection of the reviewed studies, one's personal disposition, and the too flexible nature of narrative or qualitative literature review, outcomes of such review studies have been neither consistent nor qualified to draw the sought after broad as well as strong conclusions. The present study is designed to review the available literature on subgroup of differential predictive validity. It is designed to use the quantitative approaches of literature review. Subgroups in this study have been identified on the basis of gender and race.

Differences in the predictive validity of standardized tests have been studied quite extensively by researchers in employment and educational settings. The major objective of their work has been either to reject or not to reject the hypothesis of the suitability of standardized tests for all the groups within a population.

Studies on the comparability of the predictive validity of test scores can be classified into two major models:

a) Correlational model (the differential validity model). Studies using this model usually test the null hypothesis that tests are equally valid for all subgroups. The evidence for accepting or rejecting the hypothesis is not conclusive. "The results of differential validity for minority group members (in most studies, blacks) and majority members are not consistent from one setting to another" (Linn, 1982, p. 373). On the other

hand some researchers (e.g. Reynolds, 1982; Schmidt and Hunter, 1981; Bartlet et al., 1978; Boehem, 1977) have concluded that the hypothesis of differential validity can be discredited.

b) The regression model (the differential prediction model). Students' performance is predicted by means of linear regression equations that provide estimates of the average score on the criteria measure of performance (Linn, 1982, p. 374). Studies of differential prediction have investigated the existence of significant differences between the majority and minority groups on the components of the regression equation. The null hypothesis that has been investigated states that subgroups do not differ on any of the parameters of the regression equation.

Cleary (1968) has restated the classical definition of the regression model in these words:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test is designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup (p. 115).

This model's hypothesis is rejected if significant differences between majority and minority groups are found on any of the parameters of the equation, i.e. in standard error of estimate, slopes, or intercepts. Differences in errors of estimate imply that the "magnitude of nonsystematic errors in prediction tends to be larger for one group than for members of another. The standard errors of estimate provide a more direct index of the differences in predictability between two groups than is provided by correlation coefficients" (Linn, 1982, p. 374).

Groups may differ in the degree of the steepness of their equation's slope. When the classical regression model, popularly known as Cleary's, is used, a flatter line for the minority group qualifies a fewer number of its members whenever a cut off score is used (Hymphreys, 1972, p. 64). Differences in the steepness of regression lines mean different rates of change in the criterion between the major and minor groups. Differences in the intercept when associated with equal slopes mean that the group with higher slope is under-predicted while the group with lower intercept is over-predicted, if a common regression line is used.

Two statistical procedures have been developed to test the significance of the differences in the parameters of the prediction equation for groups under comparison (Gulliksen and Wilks, 1950; Kerlinger and Pedhazar, 1973). Following Gulliksen and Wilks most of the researchers in differential prediction have tested the components of the regression model sequentially, starting with the errors of estimate. If the difference is not significant, then the slopes are compared. If their difference is not significantly different from zero, parallelism is assumed and the differences in intercepts are tested. The existence of any statistical difference in these sequential tests means that a common regression equation in not suitable.

Results of studies that have investigated differential prediction have not been conclusive, especially when blacks are compared to whites. Linn (1982) has pointed out that in the reviewed studies, more often than not, one can conclude that

the regression equations are different. On the other hand,Jensen (1980) stated that the "review of evidence on predictive bias in the prediction of black criterion performance, the bias is most commonly intercept bias and virtually always the white intercept is higher than the blacks, thereby resulting in over-prediction of blacks' criterion performance from the use of the whites' regression equation" (p. 513).

Status of literature review on differences in predictive validity in educational settings

Several major reviews have attempted to summarize the literature on either or both models of predictive validity analysis. The validity analysis. The majority of these reviews are reported in Table 1. A few reviews have not been included because: 1) they have been basically incorporated into more recent reviews by their authors (e.g., Linn, 1973 and 1975), or 2) they have not drawn conclusions nor extended the disussion beyond citations of relevant studies (e.g. Smuda, 1975).

A critical examination of these tabulated reviews shows the following points:

1) Most of these reviews have been descriptive narrations of the outcomes of primary studies. The only reviews that have tended to be more quantitative are those of Breland (1979), Cleary, Humphreys, Kendrick, and Wesman (1975), and Schmidt and Hunter (1974). But none of them has attempted to reprot integrated summary statistics on questions they addressed.

2) None of these reviews, except for Reynolds' (1982), Jensen's (1980), and Durflinger's (1943), have summarized the

studies beyond college admission.

3) All of these reviews have failed to report dissertations or other unpublished evidence, except for Educational Testing Service (ETS) or American college Testing (ACT) program reports.

4) Authors have been highly selective in the studies that have been included in their reviews.

5) No attempt has been made, except for Jensen (1980) in one instance to retrieve or reanalyze data reported by primary studies, to complete the missing information so as to fit both the correlational as well as the regression models.

6) Reviewers of studies in differntial prediction have examined the differences as reported by primary studies. As can be seen from Table 1 other predictors besides tests have been included in the prediction equations of most of the reviewed studies. It is an accepted observation that school indicators, such as high school GPA's or ranks, explain most of the variance attributed to the predictors in the regression equation that are used for admission (see Slack and Porter, 1980; Schrader, 1971). Since the subject of their review is standardized testing in particular, it seems more logical to isolate a test's contribution.

7) Many of the studies included in most of the above reviews have used samples of blacks and whites that have not come from the same college or institution. Institutional as well as racial factors may have influenced the outcomes of such studies thus throwing doubts on the plausibility of their conclusions.

The present investigation was designed to overcome these seven limitations.

Table 1

Summary of Review Studies on Tests Differential Validity and Prediction

Review Study	Number of Studies Included	Review Approach[a], (1)	Subjects' grade level[b]	Factors[c]	Predictors[d]	Setting Educational/Mixed[e]	Model[f]	Conclusions[(2)]
Linn (1982)	11	N&C	C&G	R&G	M	M	B	1) r is higher for women compared to men 2) Frequency of occurrence of higher R for white vs blacks is significant 3) LSAT[3] seems a good predictor for blacks as well as whites.
Reynold (1982)	18	N	E&S	R	M	E	B	1) Bias occurs infrequently and with no apparent observable pattern[4]
Cole (1981)	11	C	C	R	M	M	B	1) The use of combined regression equation overpredicts blacks
Scheuneman (1981)	23	N	C	R	M	M	B	1) Though tests appeared to be biased towards blacks more often than not (58% biased), the author concluded that the existence of substantial differential validity must be considered unproved.

Table 1 Continued

Jensen (1980)[5]	27	N	AL	R	M	M	B	1) Tests generally favor blacks when combined E is used.
Breland (1979)[6]	35	Q	C	R&G	M	M	B	1) r is higher for white women vs white men regardless of the indicators, but that's not the case for black women vs men. 2) Median r's are not different with race. 3) Overprediction is fairly consistent for blacks.
Cleary et al (1975)	7	Q	C	R	M	E	Reg	1) Standard tests predictions are comparable within black and white colleges. Moreover within integrated colleges the usual regression lines lead to comparable prediction for black and white students.
Schmidt & Hunter[7] (1974)	4	Q	C	R	M	M	Reg	1) Tests and E's that they have reviewed were either fair or biased in favor of blacks.
Stanley[8] (1971)	8	N	C	R	M	E	Cor	1) Regardless of the subgroup, students who are predicted to earn low grades will tend to have academic difficulties. Thus tests are valid for all subgroups.
Kendrick & Thomas (1970)	8	C	C	R&G	M	M	Cor	1) Women are found to be more predictable than men. 2) Reviewed studies indicate that tests are equally valid for blacks and whites.

Table 1 Continued

Green (1969)	12	N	C	R	M	E	Cor	1) Although studies suggest that tests predict college success equally for blacks and whites, there's a need for further research.
Levin (1965)	2	N	C	G	T	E	Cor	1) Females' R is higher than males.
Campbell (1964)	3	N	C	R	M	E	Cor	1) Tests are effective predictors of both black and white students.
Durflinger (1943)	3	N	C&S	G	T	E	Cor	1) Sex factor needs to be given a differential treatment for the purpose of predicting future performance.
Harris[9] (1931)	12	N	C	G	T	E	Cor	1) Women are more predictable than men.

---

[a] N=narrative; C=citation; Q=quasi-quantitative.

[b] E=elementary; S=secondary; C=college; G=graduate or professional school work; Al=all levels.

[c] R=race; G=gender

[d] T=tests only; M=mixed (tests and grades, for example).

[e] M=mixed; E=education.

[f] Cor=correlation model; Reg=regression model; B=both models.

Table 1 Continues

Notes:

(1) Review approaches include narration which is the traditional literature review method, citation, where studies are mainly cited in the context of the points the author is making, quantitative, where summary statistics are sought to conclude the outcomes of different studies, and "quasi-quantiative" which is the approach that reports quantitatively the outcomes of reviewed studies, without the use of summary statistics.

(2) The symbols and their meanings are:
   (1) r: simple correlation between a predictor and a criterion.
   (2) R: multiple correlation to indicate correlation between several predictors and a criterion.
   (3) E: regression equation used in prediction.

(3) Law School Admission Test.

(4) According to Reynolds (1982, p. 207), this pattern is maintained as long as tests are reliable and not specific in content. The author has been selective in studies that he has chosen to elaborate at.

(5) Most extensive review study in bias in testing in general.

(6) Most extensive review study in college admission.

(7) The analysis was conducted within the framework of two definitions of fairness. They have chosen to reanalyze four studies within this framework but discarded studies that have indicated the existence of bias.

(8) Though race is not specifically the basis for subgrouping, the reviewed studies on the disadvantaged student were mainly on blacks.

(9) Though Harris viewed race as a factor, his notion of race does not coincide with its present-day definition.

Questions of the study

The objective of this investigation is to integrate the outcomes of studies in differential test validity and prediction that have been conducted in educational settings, when gender or race is taken as a moderating factor. Specifically this study will attempt to answer the following questions:

1) Is there empirical evidence that supports the existence of differential validity when groups differ in race and/or gender?

2) Is there an under- or over-prediction of students' future performance when combined regression equations are used for subgroups that are categorized by gender or race?

Limitations of the study

1) This study is limited to integrating the statistical outcomes of studies that have investigated differential validity and prediction of academic performance.

This study does not evaluate the different models of fair test use or of item bias.

2) It does not include studies that have investigated the differential validity or prediction of non-standardized instruments or measures.

3) It is limited to studies that have used achievement tests as predictors. It does not include the outcomes of vocational or interest tests or any other non-academic measure.

4) It is limited to studies that have been conducted in

educational settings only, and it does not include employment prediction.

5) It is limited to studies that have compared black students to whites and/or females to males.

Methodology

Terminology and notation

PREDICTOR (X): Any standardized test that is used in educational settings to measure achievement or ability and has been used in the reviewed study to predict future performance.

CRITERION (Y): Any measure of the level of students' achievement (e.g. grade average) after their participation in an educational program such as undergraduate education.

MAJOR AND MINOR GROUPS: Denote males and females respectively, when gender is used as the basis of classification; denote whites and blacks respectively, when race is the basis of classification.

STUDY: A published or unpublished research report that has provided one or more units of analysis having a data set that includes, as a minimum, the groups' validity coefficients and sizes.

UNIT: Each unit provides a set of statistics of interest such as, validity coefficient (r), slope (b), intercept (a), and standard error of estimate (se), separately for mutually exclusive groups (e.g. men and women) consisting of individuals who are all members of the same schools/colleges (or group of schools/colleges). A study may contain data from several units. The number of independent units that are used in this research is

denoted as (k).

DATA-SET: A set of statistics that is obtained or retrieved from a study for each of the major and minor groups. For the correlational model each data-set is composed of two sets of sample sizes and correlation coefficients. The regression model data-sets are composed of two sets of statistics, in addition to the sample sizes and correlation coefficients. Each unit of analysis may have more than one data-set if multiple predictors or criteria are used with the same groups.

Locating studies and criteria for inclusion

To locate studies that investigated the possible differences between predictive validity values for black and white students, and for female and male students three major bibliographical sources were searched:

1) Library data bases
2) Buros' Mental Measurement Yearbook
3) Major known reviews on the predictive validity of standardized tests.

Three major library data bases were searched: Educational Resource Information Center (ERIC), Psychological Abstracts, and Dissertation Abstract International. The described bases were searched, using the specifications:

- Test (standardized or achievement or aptitude or ability),
  and
- Validity (criterion-related or predictive or subgroup or differential),
  and
- Race (anglo or white or black or negro or disadvantaged or minority) or cultural fair or bias or sex.

This search resulted in 806 titles in ERIC, 212 in Psychological Abstracts, and 208 in Dissertation Abastracts International.

Bibliographies of the Scholastic Aptitude Test (SAT), American College Testing Program (ACT), and the Cooperative School and College Ability Test (SCAT) in the different editions of Buros' Mental Measurement Yearbook and Tests in Print 11 were searched.

The major reviews that were located and searched were: Wagner (1934), Harris (1940; 1935), Durflinger (1943), Lavin (1965), Shuey (1966), Dreger and Miller (1968), and Samuda (1975).

The abstracts, when available, or located materials themselves, were reviewed to judge their suitability for the present investigation. Each selected study met the following criteria:

1) Used standardized achievement tests as predictors, including tests of general achievement and ability.

2) Reported grade performance or scores obtained on standardized tests as criteria.

3) Reported the correlation and/or regression coefficients by race and/or by gender.

4) Reported subgroup sizes.

5) Reported that at least 80% of the minority subgroups were black students (when race was the basis of classification).

6) Indicated that a period of at least eight weeks separated the administrations of the predictor and the criterion (to ensure the predictive function of the test).

7) Indicated that students of both subgroups, whether blacks and whites and/or females and males, came from the same schools/colleges or groups of schools/colleges (to control

institutions as a contaminating factor that could account for differential validity).

Authors of the studies that could qualify for this investigation, except for some missing data such as size of the group, zero-order coefficients, or necessary descriptive statistics, were contacted by mail when it was feasible to do so. Unfortunately, only one of the ten authors contacted responded positively.

The sample

After eliminating all studies that did not meet the prescribed criteria, 40 studies that reported relevant data-sets for black and white differences and 53 studies that reported separate data-sets for males and females were located. Several studies contributed more than one unit of analysis.

Most of the located studies that were judged to be suitable for the present integration have investigated the differential validity and prediction at the college level(80%-85%); a few have studied predictors that are used in elementary and middle school, and the number of studies that investigated high school prediction is very small.

The following procedures were taken in processing the obtained data-sets. Most of them were meant to ensure the independence of the different data-sets that were used in the study synthesis.

(1) Except for point (2) below, data-sets reported by different studies were treated as units of analysis, as long as they are based on independent samples.

(2) When data in a study were presented for the four race-sex combinations, two separate units were used in the synthesis, one for the synthesis by race and one for synthesis by gender. The rationale for this decision was that independence of units is preserved within each integration study.

(3) Some studies reported more than one data-set for the different components of the predictor (e.g., SAT-Verbal and SAT-Math) or for more than one predictor for the same group, and no precedence could have been established. In such situations, the outcomes from each analysis were averaged. Only those averages were used to represent such units in later analysis.

If the relationships of several predictors with the same criterion were investigated, the data-set depicting the conventional predictors-criterion relationships was selected for this study. For example, if a study reports the correlation of scores on the Iowa Test of Basic Skills (ITBS), given at the eighth grade, with first year GPA, as well as the ACT with the same criterion, the later data-set was chosen.

(4) Some studies have reported more than one data-set for the same groups, where one set depicts the correlation between a certain predictor and students' grades, while the other depicts the correlation of the same predictor but with students' ranks in their classes. In such situations the first data-set was chosen. The rationale for this decision was that since the majority of studies reported students' grades rather than their ranks, therefore the above choice will increase the homogeneity of the

Table 2

Number of Published and Unpublished
Studies and Units in Differential Validity and Prediction

	Differential Prediction				Differential Validity			
	Females/Males		Blacks/Whites		Females/Males		Blacks/Whites	
	Study	Unit	Study	Unit	Study	Unit	Study	Unit
Journals/Books	20	57	14	33	32	90	20	39
Reports	9	58	6	18	11	97	6	19
Unpublished Papers & Dissertations	9	32	13	39	10	33	14	40
Total	38	147	32	90	53	220	40	98

criteria.

(5) Some studies reported more than one data-set based on the same group to capture longitudinal changes in the predictive validity of some instruments. In such a case, the data-set with the shortest whole unit of time in the educational setting, such as academic year, was chosen. For example, if the Scholastic Aptitude Test (SAT) is correlated with first year GPA, and final year GPA, the first data-set was used; or if SAT is correlated with first semester as well as freshmen year GPA, the later data-set was chosen.

Preliminary data analysis

The differences between major and minor groups on the following statistics were used in the present research:

(1) The validity coefficient (r) between the predictor score and the criterion value.

(2) The regression coefficient (b).

(3) The intercept (a).

(4) The standard error of estimate (se).

Most of the studies that were located reported the r; very few reported the regression statistics and, if so, they reported the partialled coefficients and multiple R's, as they used more than one predictor. On the basis of the available descriptive statistics, the researchers retrieved the groups' regression coefficients and standard error of estimates, and the outcomes of the conventional statistical tests of significance that are used to compare each pair of relevant statistics for major and minor groups.

Meta-analytical approaches used in the analysis

The outcomes of the studies used in this investigation were integrated by several approaches. They are grouped in this section into four categories.

1.0 Integrating effect size (see Glass, 1976):

Let E be a statistic of interest for group g. The effect size for an individual unit of analysis based on that statistic is given by

$$d = ({}_{1}E_{i} - {}_{2}E_{i})/ s_{i} \text{ , where } i = 1,\ldots,k \qquad (1)$$

and where s_i is a standardizing statistic and k is the number of units. The effect size of all studies is some composite mean of each unit d_i, namely $\bar{d}$.

It was decided to base the standardizing statistic, s_i, on pooling both groups standard deviations, according to the size of each group as in equation (2).

$$s_i = \text{SQRT} \left[\frac{df_{1i}\,({}_{1}s_{i})^2 + df_{2i}\,({}_{2}s_{i})^2}{df_{1i} + df_{2i}} \right] \qquad (2)$$

where, ${}_{g}s_{i}$ = standard deviation of group g, and ${}_{g}df_{i}$ = df of group g in unit i.

This measure was adapted for several reasons. First, the studies involved in this research are non-experimental and consequently, no control groups were involved. Therefore the s suggested by Glass (1976) was not appropriate. Second, the individuals in each unit come from the same school/ college, therefore it is more natural to assume the homogeniety of their variance. Third, in most studies the difference in the sizes of

major and minor groups is substantial. Consequently, the variance of the larger group tends to be more stable and accurate estimator of the population variance, assuming other factors are equal. Finally, Hedges has observed that the d_i that is based on the weighted s_i is a less biased estimater of the true effect size, than that based on the control groups s (1981, p. 110).

1.1 Differences in validity coefficients: The effect size for the differences in validity coeeficients is given by equation 1, where, E_i is the Fisher's z transformed value of r. In this case the s_i is unity, since r itself is a standard unit.

1.2 Differences in slopes: The effect size for the differences in slopes is given by equation 1, where E_{gi} is the observed regression coefficient for group g in unit i, the s_i is the pooled standard error of the slopes of major and minor groups computed from equation 2.

1.3 Differences in intercepts: The effect size for the differences in intercepts is given by equation 1, where E_{gi} is the observed regression intercept for the equation computed on students in group g in unit i, and s_i is the pooled standard deviation of the criterion y, computed from equation 2, with df equal to (n - 1).

1.4 Differences in standard errors of estimate: The effect size for the standard errors of estimate is given by equation 1, where E_{gi} is the observed se for group g in unit i, and s_i is the pooled standard error or estimate for the major and minor group as given by

$$s_i = \left[\frac{{}^{g}SS_{resid_i} + {}^{g}SS_{resid_i}}{{}^{g}df_i + {}^{g}df_i} \right]^{\frac{1}{2}} \qquad (3)$$

Each unit of analysis contributed one d_i , which can be the only effect size obtained from that unit, or the average of several effect sizes that are based on the same group in the unit. The sample average of their d_i's, $\bar{d}$, is derived and tested against the null hypothesis the true effect size equal 0. The standard error of $\bar{d}$ is given by

$$s_{\bar{d}} = \frac{\{\sum (d_i - \bar{d})^2 / (k-1)\}^{\frac{1}{2}}}{\sqrt{k}} \qquad (4)$$

The outcome of $d/ s_{\bar{d}}$ is compared to the critical value at the chosen alpha level, using Student's t distribution with df = (k-1). The rationale for this test are: (1) since we have a fairly large number of units, we can safely invoke the central limit theorem. This rationale persists as long as k is large. (2) Hedges has observed that the asymptotic distribution of large samples ($_1n_i$, $_2n_i$> 10) provides a satisfactory approximation to the normal distribution (Hedges, 1982, p. 13 - 28).

2.0 Integrating correlations coefficients:

Glass et al. (1981) have pointed out the suitability of correlation coefficients and proportions as tools that can be used in quantitative study synthesis as alternatives to the effect size. Since the correlation coefficient is the major measure that has been used by studies that have investigated the issue of subgroup validity of standardized tests, it therefore seems appropriate to employ methods that are based on this coefficient in this investigation.

2.1 Averaging the weighted differences between each pair of validity coefficients. Each r is converted into Fisher's z, and the difference between the pair of z's was pooled to derive the

average as given by

$$\bar{d}_z = \frac{\sum d_{z_i} w_i}{\sum w_i} \quad , i = 1,\ldots,k \qquad (5)$$

where d = the difference between a pair of Fisher z transformed correlation coefficients, and

$$w = 1/ \left[\frac{1}{({}_1n_i - 3)} + \frac{1}{({}_2n_i - 3)} \right] \qquad (6)$$

where ${}_1n_i$, ${}_2n_i$ are the major and minor groups size in the unit of analysis.

The standard error of the $\bar{d}_z$ is given by

$$s_{\bar{d}_z} = \left[\frac{1}{\sum w_i} \right]^{\frac{1}{2}} \qquad (7)$$

The significance of the averaged z was tested as given by (8) using the normal distribution, since the sampling distribution of $({}_1z_i - {}_2z_i)$ is normal.

$$d_z / s_{\bar{d}_z} = z \qquad (8)$$

2.2 Simple average of validity coefficients: Validity coefficients for both major and minor groups in each unit were averaged. The difference in averages was assessessed by reference to the t-distribution.

Results and analysis

Question one: Is there empirical evidence that supports the existence of differential validity when groups differ in race and/or in gender?

Rephrasing this question, this research tested the following hypothesis:

$$Ho: \quad P_j - P_n = 0$$

where j denotes major group, n denotes minor group and, P denotes population validity coefficient.

The obtained evidence was tested using three different methods.

(1) Simple Average of Validity Coefficient (see method 2.2 above).

The averages of validity coefficients, r's were .316, .371 for blacks and whites repectively, and .420, and .353 for females and males (see Table 3). Considering the size of the aggregated samples these correlations, as well as their differences ($t(97) = 3.60$, $p = .0004$) and ($t(219) = -7.611$, $p = .000$) for blacks and females units respectively, are significant. But the differences in terms of magnitude are modest.

(2) Effect Size (see method 1.1 above).

Each pair of validity coefficients was converted into Fisher's z scores. The averaged differences by units were averaged, then t-tested by Students's t test for dependent samples (see Table 4). The average differences in z units in blacks' and females' studies were .066 and -.080 respectively. Considering the number of integrated studies, differences were significant for both factors ($t(97) = 3.625$, $p = .0005$; $t(219) = -7.697$, $p = .000$).

(3) Averaging Weighted Differences (see method 2.1 above).

This approach weighs the difference obtained from each pair of z-transformed r's by the reciprocal of their variance. So studies with smaller variances and more consistent outcomes would

Table 3

Descriptive Values of Tests Validity Coefficients of Different Subgroups

Subgroup	Number of students	Number of units	Average r	Standard Deviation[1]	Minimum value	Maximun value	t value	p-value
By race								
Whites	79144	98	.371	.159	.105	.820	23.049	.000
Blacks	9932	98	.316	.162	-.200	.730	19.284	.000
By gender								
Males	58773	220	.353	.168	-.366	.735	31.219	.000
Females	54664	220	.420	.143	-.092	.820	43.497	.000

[1] Standard deviation of the observed validity coefficients within each subgroup.

Table 4

Outcomes of Effect Size Differences
Between Observed Validity Coefficients

Average d_r [1]	Number of Units	Standard Error of d_r	t value	p-value	Average d_z [2]	Standard Error of d_z	t value	p-value
Whites-Blacks Differences								
.0545	98	.0150	3.650	.0004	.0660	.0182	3.624	.0005
Males-Females Differences								
-.0675	220	.0089	-7.611	.0000	-.0800	.0104	-7.697	.0000

[1] Using validity coefficients.

[2] Using transformed r's into Fisher's z.

Table 5

Differences in Validity Coefficients Using Weighted Averaged Differences Approach [1]

Factor	Number of Students	Number of Units	Average of Differences	Standard Error of Average	Standard Deviate Scores	p-value
Race	89076	98	.0403	.0118	3.472	.0005
Gender	113437	220	-.0595	.0061	-9.716	.0000

[1] See equations 5 to 8

have larger weight. The outcomes of this method (see Table 5) showed that the average weighted differences were .04 and -.06 from the blacks-whites groups, and females-males group respectively. Both of these differences were statistically significant, though they were small in magnitude ($t(97) = 3.427$, $p = .0005$ for blacks; $t(219) = -9.716$, $p = .000$ for females).

In summary, these analyses indicated that on the average validity coefficients for blacks were lower than those for whites. Moreover, the validity coefficients for females were on the average, higher than those for males.

These analyses indicated that the magnitude of the differences, though statistically significant, was too small to be meaningful. The importance of such differences comes from their persistence. In 65.3% of the units the validity coefficients were higher in favor of whites ($z = 2.873$, $p < .0021$), and in 72.6% of the units they were in favor of females ($z = 6.348$, $p < .0 38$).

These findings are somewhat consistent with the outcomes of other studies (e.g., Linn, 1982). Most of the investigators in the area of differential validity have attributed these differences, especially for blacks, to design factors such as small sample size and range restriction. Only 13% of the units that considered race as a factor had 29 or fewer subjects in the smaller of the two samples, while 12.3% of the units were considered restricted by the criterion the ratio of minor group variance to major group variance, on the predictor is .70 or less. These observed characteristics of

the units under consideration throw some doubts on the plausability of the above explanations.

There is no doubt that the mean score of most of black students' samples on the predictor falls approximately one standard deviation below that of the whites'. In this investigation, the difference between the means of whites and blacks on each predictor and criterion was standardized, using the black students' standard deviations on each measure respectively, i.e.(($_{w}\bar{X}_i - {}_{b}\bar{X}_i$) / $_{b}SD_{X_i} = d_i$). The average of the 230 observations' d_i was 1.215. This difference is illustrated in figure 1, that shows almost 89% of the black students scored below the white students' mean in their relevant predictors. On the other hand, the average standardized difference for the same groups on the relevant criteria was .825 in favor of whites and about 79% of black students scored below the white students' relevant criteria. Thus, although whites' mean score superiority is maintained whether predictors or criteria are considered, their rank superiority droppped about 10% on the criterion score.

It should be noted that this relative improvement in performance for the blacks was not in contradiction with the observation that differences in tests predictive validity coefficients were small in magnitude. The obtained validity coefficients were a function of the relative position of individuals within each subgroup. Therefore, the improvement of blacks' performance on the criteria as a group did not necessarily mean variations within the order of black students in their respective subgroups.

The differences between males' and females' mean score on each predictor and criterion in 370 data-sets were standardized using the females standard deviation on each measure. The average difference on the predictors was .172 in favor of males. That is, the score of the males was on the average, higher by about .17 standard deviation than that of the females on the predictors. In contrast, females' mean score was on the average, higher than that of the males' by .326 of females' criterion standard deviation. 56.8% of the females scored below the mean of the male students on the predictor, while 62.9% of the males scored below the mean of female students on the criterion.

This discrepancy has been investigated by many researchers (e.g., Gamache and Novick, 1983). It has been attributed to many factors such as the subjectivity of the criteria as measures of students'performance, and the types of major that attract the females relative to males.

Question Two: Is there an under- or over-prediction of students' future performance when combined regression equations are used when groups are categorized by gender or race?

Overprediction occurs when the regression line of the minor group, for example, falls below that of the major, group, assuming that both lines are parallel. In contrast, prediction occurs when the minor group's regression line falls above the major's, again assuming parallelism.

To translate the second question in terms of the above elaboration, this study investigated the following questions:

i) are the standard errors of estimate (se),

ii) are the slopes, i.e., the regression coefficients (b), and iii) are the intercepts, i.e., the constant regression coefficients, (a), different for each group?

The outcomes are summerized in Table 7. The results indicated the following:

(1) Small but significant differences appear in standard errors if estimate when race or gender were considered as factors.

(2) Due to the small size of differences in standard errors of estimate in both comparison, the analysis for slopes was carried out. Significant differences appear in slopes when race or gender were considered as factors. Regression coeeficients were higher for whites compared to blacks, but they were higher for females compared to males. If such outcome is true, it means that more females and more whites qualify for selection, for admission, or for similar purposes than do males and blacks, all things being equal.

(3) Though significant differences in slope would rule out the necessity of intercept comparison (Gulliksen and Wilks, 1950) the authors integrated the differences in intercepts on both race and gender. The results indicated that differences in intercepts were not significant.

The outcomes of the slope comparisons motivated the authors to examine the regression co-efficents at the unit level for both race and gender. The results indicated the following:

1) Slopes are equal in 70 out of 90 units when blacks and whites were compared, and in 119 out of 147 units when gender was the comparison factor. Both of the above proportions are

Table 6

Summary of the Outcomes of the

Meta - Analytical Tests

Factor		Standard eror	Slope	Intercept	No. of Units
	M	0.048	1.208	0.006	90
Race	SD	0.158	3.301	0.117	
	t	2.850*	3.470*	0.050	
	M	0.067	-0.795	0.204	147
Gender	SD	0.117	1.868	2.307	
	t	6.943*	-5.162*	1.070	

* ($P < .01$)

significant (P=.000).

2) Wherever differences in slopes were not established, over- or under-prediction was not constant.

In 41% of the parallel black/white units, intercepts were higher for Blacks (P=.09). 53% of the female/male units were higher for females (P=.23).

Given the above findings that are related to question two of this study, one can conclude the differences in regression equations between each two groups are expected to exist, though in small magnitudes. And it seems that evidence indicated that, even with such differences, the claim of constant over-prediction, or under-prediction in favor of one particular minor or major group, regardless of the factor, was not warranted and should be rejected.

Conclusions

This investigation reviewed the literature on the differential validity and prediction of standardized tests for groups that were divided on the basis of race (specifically, blacks in relation to whites) and of gender. The study employed some of the techniques in quantitative study integration. The retrieved data-sets were averaged into 147 and 90 independent units for females-males, and blacks-whites comparisons respectively.

The results of integrating validity coefficients were:

1) The averaged validity coefficients for whites were higher than those for blacks.

2) The averaged validity coefficients for females were higher

than those for males.

3) The averaged differences in validity coefficients for black units as compared with white units and for female units as compared with male units were significant, though the averaged differences were too small in magnitude to be meaningful.

When the slopes were integrated, it was found that:

1) The averaged regression coefficients for whites were higher than those for blacks. This indicated steeper slopes for whites, thus ensuring more white candidates whenever a selection decision was to be made.

2) The averaged regression coefficients for males were higher than those for females. This indicated steeper slopes for males, thus ensuring more male candidates whenever a selection decision was to be made.

But further analysis has shown that:

1) The regression lines in about 78% of the integrated black/white units were parallel.

2) The regression lines in about 81% of the integrated female/male units were parallel.

3) Wherever equality of slopes was established, over- or under-prediction was not constant. In 41% of the parallel black/white units, intercepts were higher for blacks ($p = .09$). 53% of the female/male units were higher for females ($p = .23$).

It can be concluded as a result of this investigation that tests as predictors of future performance are as "good" or "bad" predictors for blacks and females as they are for whites and males. Since the studies in predictive validity are sample-

dependent as is the case in social sciences in general, variability in outcomes is likely to be observed and small differences in the results of individual validity studies will always persist.

Blacks are more likely to score below the mean of the whites students on tests that are used for prediction, especially in college admission tests. The magnitude of that difference is reduced on the criterion measures, e.g. GPA. This improvement of black students' performance on the predicted measure relative to the whites' could provide the justification for the use of separate admission criteria for both groups. But this is a social policy issue, and it depends heavily upon one's own conceptualization of the question of fairness. The specification of the selection criteria is not within the objectives of this investigation. Such criteria are likely to be dependent upon the the characteristics of the institutions' local social settings.

Females' mean scores on the standardized tests, when used as predictors, tended to fall below those of the males. On the other hand, females' mean scores on the average were higher than males' when the two groups were compared on the criteria. This variation in females' performance on the predictors and the criteria relative to the males' has been noted by earlier researchers (e.g. Gamache and Novick, 1983).

Finally, the results of this investigation should be considered within its limitations. This study did not go beyond integrating the outcomes of the relationship of one predictor with one criterion at a time. In actual academic situations,

selection is based on more than one predictor. Therefore, these outcomes should not be extended for the purpose of policy planning beyond the limited context of this investigation.

List of Studies Included in The Synthesis

Abelson, R. P. (1952). Sex differences in predictability of college grades. Educational And Psychological Measurement, 12, 638-644. (code 24).

Astin, A. W. (1971). Predicting academic performance in college selectivity data for 2300 American colleges. New York : Free Press. (code 66).

Ayres, Q. W. (1983). Student achievement at predominantly white and predominantly black universities. American Educational Research Journal, 20, 291-304. (code 100).

Baggaley, A. R. (1974). Academic prediction at an ivy league college, moderated by demographic factors. Measurement and Evaluation in Guidance, 6, 232-235. (code 12).

Boney, J. D. (1966). Predicting the academic achievement of secondary school negro students. Personnel and Guidance Journal, 44, 700-703. (code 98).

Baird, L. L. (1969). The prediction of grades in occupational and academic curricula in two-year colleges. Journal of Educational Measurement, 6, 247-254. (code 96)

Bowers, J. (1970). The comparison of GPA regression equations for regularly admitted and disadvantaged freshmen at the University of Illinois. Journal of Educational Measurement, 7, 219-223. (code 69).

Breland, H. M., & Griswold, P. A. (1982). Use of a performance test as a criterion in a differential validity study. Journal of Educational Psychology, 74, 713-720. (code 11).

Breland, H. M. (1978, March). Sex and ethnic comparisons of test validity using a blind-scored essay as a criterion. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada. (code 42).

Brown, J. L. B., & Lightsey, R. (1970). Differential predictive validity of SAT scores for freshmen college english. Educational and Psychological Measurement, 30, 961-965. (code 74).

Casserly, P. L. (1982). Older students and the SAT (Report No. 82-8). New York: College Entrance Examination Board. (ERIC Document Reproduction Services No. 226 051). (Code 37).

Chase, L. I., & Thompson, R. (1973). Predicting success for university freshmen (Report No. 1, Supplement 4). Bloomington: Indiana Studies in Prediction, Bureau of Educational Studies and Testing, Indiana University. (ERIC Document Reproduction Services No. 077 488). (code 51).

Cherdack A. N. (1971). The predictive validity of the Scholastic Aptitude Test for disadvantaged college students enrolled in a special education program. Los Angles: University of California. (ERIC Document Reproduction Services ED 051 759). (code 60).

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124. (code 92).

Crisco, J. J. (1975). The prediction of academic performance for minority engineering students from selected academic proficiency, personality, cognitive style, and demographic variables. Dissertation Abstracts International, 36, 6545A. (University Microfilm No. 76-8634). (code 87).

Davis, J. A., & Temp, G. (1971). Is SAT biased against black students? College Board Review, 81, 4-9. (code 25)

Davis, J. A., & Kerner, S. E. (1971). The validity of tests and achievement in high school for predicting performance in the public universities of North Carolina with special attention to black students (RIS Report No. 71-3). Princeton, NJ: Educational Testing Services. (ERIC Document Reproduction Services No. ED 065 536). (code 25).

Davis, S. C., Loeb, J. W., & Robinson, L. (1970). A comparison of characteristics of negro and white college freshmen classmates. The Journal of Negro Education, 39, 359-366. (code 34).

Durio, H. F., Kildow, C. A., & Slover, J. A. T. (1980, April). Ethnicity and sex differences in use of college entrance examination, mathematics achievement, and high school rank as predictors of performance and retention among engineering students. Paper presented at the annual meeting of the American Educational Research Association, Boston. (code 36).

Goldman, R. D., & Hartig, L. K. (1976). The WISC may not be a valid predictor of school performance for primary-grade minority children. American Journal of Mental Deficiency, 80, 583-587. (code 48).

Goldman, R., & Hewitt B. N. (1976). Predicting the success of black, chicano, oriental, and white college students. Journal of Educational Measurement, 13, 107-117. (code 7).

Goolsby, T. M., & Frary, R. B. (1970). Validity of the Metropolitan Readiness Test for white and negro students in a southern city. Educational and Pyschological Measurement, 30, 443-445. (code 39)

Green, R. L., & Farquhar, W. W. (1965). Negro academic motivation and scholastic achievement. Journal of Educational Psychology, 56, 241-243. (code 97).

Gruen, R. S. (1971). Prediction of end-of-year reading achievement for first and third grade pupils. (Doctoral dissertation, The Pennsylvania State University). Dissertation Abstracts International, 32, 6198A. (code 63).

Hammond, L. R., & Rosick, D. C. (1971, February). Problems of prediction with college education achievement project at Allen University. Paper presented at the annual meeting of the American educational Research Association, New York. (code 57).

Hills, J., & Stanely, J. C. (1970). Easier test improves prediction of black students' college grades. Journal of Negro Education, 39, 320-324. (code 33).

Hills, R., Emory, L. B., & Masters, P. B. (1962). Freshmen norms for the university system of Georgia, 1960-1961. Atlanta: Regents of the Univsetrsity System of Georgia. (code 103).

Hodge, C. C. Jr. (1981). Predictive study of the Meeting Street School Screen Test and academic achievement from grades two through six by race, sex, kindergarten experience. Dissertation Abstracts International, 42, 3118A. (University Microfilm No. 81-29498). (code 88).

Jackson, R. A (1955). Prediction of the academic success of college freshmen. The Journal of Educational Psychology, 46, 96-301. (code 46).

Jacobs, J. N. (1959). Aptitude and achievement measures in predicting high school academic success. Personnel and Guidance Journal, 37, 334-341. (code 27).

Jones, D. J. (1970). Correlates of academic success in a predominantly black, open-door, public, urban community collge. (ERIC Reproduction Services No. ED 039 877). (code 67).

Jones, E. S. (1930). Studies from the office of personnel research. The University of Buffalo Studies, 8(1), 1-87. (code 53).

Jones, W. P. (1968). Predicting academic performance with the Short Test of Educational Ability considering social class and sex differences. Dissertation Abstracts, 29, 3832A. (University Microfilms No. 69-0793). (code 91).

Kallingal, A. (1971). The prediction of grades for black and white students at Michigan State University. Journal of Educational Measurement, 8, 263-265. (code 94).

Klugh, H. E., & Bierley, R. (1959). The school and college ability test and high school grades as predictors of college achievement. Education and Psychological Measurement, 19, 625-626. (code 26).

Larson, J. R., & Scontrino, P. (1976). The consistency of high school grade point average and mathematical portions of the Scholastic Aptitude Test of College Entrance Examination Board, as predictors of college performance: An eight year study. Educational and Psychological Measurement, 36, 439-443. (code 38).

Layton, W. L., & Swanson, E. O. (1958). Relationship of ninth grade Differential Aptitude Test scores to eleventh grade test scores and high school rank. Journal of Educational Psychology, 49, 153-155. (code 29).

Lins, L. J., Abell, A. P., & Hutchins, H. C. (1966). Relative usefulness in predicting academic success of the ACT, SAT, and some other variables. The Journal of Experimental Education, 35(2), 1-29. (code 71).

Loyd, B. H., Forsyth, R., & Hoover, H. D. (1980, April). Relationship of elementary and secondary achievement test scores to later academic success. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (code 72).

Lunemann, A. L. (1972). The Predictive Validity of IQ tests for different ethnic groups in a desegregated elementary school setting (Doctoral dissertation, University of California, Berkeley). American Doctoral Dissertations, 11, 146. (code 59).

Magoon, R. A. (1967). The prediction of feshmen academic performance in an urban southern college. Dissertation Abstracts, 28, 3371A. (University Microfilm No. 68-3140). (code 82).

Mathis, H. R. (1972). The validity of the American College Test in the prediction of academic achievement in community colleges of Appalachian Kentucky. (Doctoral dissertation, North Carolina State University, Raleigh). Dissertation Abstracts International, 33,3391A. (code 81)

McCornack, R. L. (1983). Bias in the validity of predicted college grades in four ethnic minority groups. Educational and Psychological Measurement, 43, 517-522. (code 93).

McKelpin, J.P. (1965). Some implications of the intellectual characteristics of freshmen entering a liberal arts college. Journal of Educational Measurement, 2, 161-166. (code 95).

Merrit, R. (1972). The predictive validtiy of the American College Test for students from low socioeconomic levels. Educational and Psychological Measurement, 32, 443-445. (code 4).

Michael, W. B., Cathcart, R., Zimmerman, W. S., & Milfs, M. (1963). Gains in various measures of communication skills relative to three curricular patterns in college. Educational and Psychological Measurement, 23, 365-374. (code 9).

Michael, W. B., & Jones, R. A. (1963). Stability of predictive validities of high school grades on Scholastic Aptitude Test of the College Entrance Examination Board for liberal arts students. Educational and Psychological Measurement, 23, 375-378. (code 8).

Mitchel, B. C. (1967). Predictive validity of Metropolitan Readiness Tests and the Murphy-Durrell Reading Readiness analysis for white and negro pupils. Euducational and Pyschological Measurement, 27, 1047-1054. (code 6).

Nagle, R. J. (1979). The predictive validity of the Metropolitan Readiness Tests, 1976 edition. Educational and Psychological Measurement, 39, 1043-1045. (code 16).

Oakland, T. (1978). Predictive validity of readiness tests for middle and lower socioeconomic status Anglo, Black and Mexican American. Journal of Educational Psychology, 70, 574-582. (code 20).

Oakland, T. (1980, September). Predictive validity for the WISC-R IQs and estimated learning potential. Paper presented at the annual meeting of the American Psychological Association, Montreal. (ERIC Reprodiction Services No. ED 194 546). (code 55).

Pandey, R. E. (1971). The SCAT and the race. Psychological Reports, 28, 459-462. (code 52).

Perry, D. K., Joslyn, E. G., & Swanson, E. O. (1971). Predicting college grades of Minnesota high school students. (OSA Research Bulletin, Vol 7, No. 7). University of Minnesota, Research Bulletin of the Office for Students Affairs. (ERIC Document Reproduction Services No. ED 055 105). (code 58).

Perry, L. A. (1974). A comparative study of the predictive validity of two reading readiness tests. Dissertation Abstracts International, 40, 3975A. (University Microfilms No. 80-01302). (code 86).

Pfeifer, C. M. Jr., & Sedlacek, W. E. (1971). The validity of academic predictors for black and white students at a predominantly white university. Journal of Educational Measurement, 8, 25-261. (code 21).

Pfeifer, C. M. (1972). The academic ability and university climate in biracial academic predictors (Doctoral Dissertation, University of Maryland). Dissertation Abstracts International, 34, 905B. (code 84).

Pitcher, B. (1977). Subgroup validity study. Reports of LSAC sponsored research: Vol. III, 1975-1977 (Report No. LSAC-76-6). Law School Admission Council, Princeton, NJ. (code 110).

Pitcher, B. (1976). Predicting Law school grades for female law students. Reports of LSAC Sponsored Research: Vol II, 1970-1974 (Report No. LSAC-74-3). Law School Admission Council, Princeton, NJ. (code 62).

Reschly, D. J. & Sabers, D. L. (1979). Analysis of test bias in four groups with the regression definition. Journal of Educational Measurement, 16, 1-9. (code 10).

Reynolds, C. R. (1980). An examination of bias in a preschool test battery across race and sex. Journal of Educational Measurement, 17, 137-146. (code 5).

Russel, H., Michael, W. B., & Martois, H. (1976). The prediction of success of three ethnic groups in the academic components of a nursing-training program at a large metropolitan hospital. Educational and Psychological Measurement, 36, 421-431. (code 78).

Sample, D. D., & Seymour, W. R. (1971). The academic success of black students: A dilemma. The Journal of College Student Personnel, 12, 243-247. (code 61).

Schrader, W. B., & Pitcher, B. (1974). Predicting law school grades for Black American law students. Reports of LSAC sponsored research: Vol II, 1970-1974 (Report No. LSAC-73-6). Law School Admission Council, Princeton, NJ. (code 102).

Scott, C. C. (1976). Longer-term predictive validity of college admission tests for Anglo, Black, and Mexican American students. Dissertation Abstracts International, 35, 2733A. (University Microfilm No. 76-25697). (code 85).

Sedlacek, W. E., & Brooks, G. C. Jr. (1972). Predictors of academic success for university students in special programs. (CSC Report No. 4-72) College Park: University of Maryland, Cultural Study Center. (ERIC Document Reproduction Services No. ED 073 222). (code 41).

Sewell, T.E. (1979). Intelligence and learning tasks as predictors of scholastic achievement in black and white first grade children. Journal of School Psychology, 17, 325-332. (code 65).

Siegelman, M. (1971). SAT and high school average predictions of four year college achievement. Educational & Psychological Measurement, 31, 947-950. (code 73).

Stroup, A. L. (1970). The prediction of academic performance from personality and aptitude variables. The Jounal of Experimental Education, 38 (3), 83-86. (code 77).

Thomas, C. L. (1971). The relative effectiveness of high school grades and standardized test scores for predicting college grades of black students. (Doctoral dissertation, Johns Hopkins University). Dissertation Abstracts International, 32, 2495A. (code 64).

Thomas, C. L. (1973, February). The overprediction phenomenon among black correlations: Some preliminary consideration. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (code 50).

University of Rochester (1984). [Predictors and first year grade point average for classes of 1984 to 1987]. Unpublished data. (code 101).

Wagner, M. E. (1934). Prediction of college performance. University of Buffalo Studies, 9, 125-144. (code 68).

Wagner, M.E. & Strabel, E. (1935). Homogeneous grouping as a means of improving the prediction of academic performance. Journal of Applied Psychology, 19, 426-446. (code 35).

Wallace, P. E. (1979). The efficacy of tenth grade DAT scores in predicting scores on the SAT. Dissertation Abstracts International, 40, 4418A. (University Microfilms No. 80-04843). (Code 83).

Wilhelm, E. M. (1971). Predicting academic achievement at a multi-purpose college. Dissertation Abstracts International, 32, 4261A. (University Microfilms No. 72-7198). (code 80).

Wilson, K. M. (1978). Predicting the long-term performance in college of minority and nonminority students: A comparative analysis in two collegiate settings. (Reoprt No. RDR 77-78, 3). New York: College Entrance Examination Board. (code 75).

Wolfson, S. J. C. (1971). A comparative study of the prediction of academic achievement of Anglo-American, Mexican-American, and Negro students in junior high school. Dissertation Abstracts International, 33, 1032A. (University Microfilms No. 72-7198). (code 90).

Wright, R. J., & Bean, A. G. (1974). The influence of socioeconomic status on the predictability of college performance. Journal of Educational Measurement, 11, 277-84. (code 3).

Young, G. E. (1972). Differential predictive validity of the American College Test (ACT) for minority and nonminority students. Dissertation Abstracts International, 33, 1453A. (University Microfilms No. 72-27143). (code 89).

LIST OF REFERENCES

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, P. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. Personnel Psychology, 31, 233-241.

Berk, R. A. (Ed.). (1982). Handbook of methods for detecting test bias. Baltimore: The Johns Hopkins University Press.

Boehm, V. R. (1977). Differential prediction: A methodological artifact? Journal of Applied Psychology, 62, 146-154.

Bruning, J. L. & Kintz, B. L. (1968). Computational handbook of statistics. Glenview, Il.: Scott, Foresman.

Buros, O. K. (Ed.). (1978). The eighth Mental Measurement Yearbook (Vols. 1-2). Highland Park, NJ: Gryphon Press.

Breland, H. M. (1979). Population validity and college entrance measures (Research Monograph No. 8). New York: The College Board.

Campbell, J. (1964). Testing of culturally different groups (Research Bulletin 64-34). Princeton, NJ: Educational Testing Service.

Callender, J. C., & Osburn, H. G. (1980). Development and test of a new model for generalization of validity. Journal of Applied Psychology, 65, 543-558.

Clark, K. B. & Plotkin, L. (1963). The integrated colleges. New York: National Scholarship Services and Fund for Negro Students.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., and Wesman, A. (1975). Educational uses of tests with disadvantaged students. American Psychologist, 30, 15-41.

Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115-124.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Cole, N. S. (1981). Bias in testing. American Psychologist, 36, 1067-1077.

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. Review of Educational Research, 52, 291-302.

Darlington, R. B. (1982). Notes on meta-analysis. Unpublished manuscript, Cornell University, Ithaca, NY.

Darlington, R. B. (1975). Radicals and squares. Ithaca, NY: Logon Hill Press.

Durflinger, G. W. (1943). The prediction of college success- a summary of recent findings. Journal of The American Association of Collegiate Registrars, 19(1), 68-78.

Dreger, R. M. & Miller, K. S. (1968). Comparative psychological studies of negroes and whites in the United States: 1959-1965. Psychological Bulletin Monograph Supplement, 70, (3,pt. 2).

Eysenck, H. J. (1978). An exercise in mega-silliness. American Psychologist, 33, 517.

Fishman, J. A. & Pasanella, K. (1960). College admission-selection studies. Review of Educational Research, 30, 298-310.

Fincher, C. (1977). Standardized tests, group differences, and public policy. The College Board Review, 103, 19-31.

Fisher, R. A. (1948). Statistical methods for research workers (10th ed.). Edinburgh: Oliver and Boyd.

Gemache, L. M. & Novick, M. R. (1983). Choice of variables and gender differentiated prediction within selected academic programs (Research Report No. 105). Iowa: The University of Iowa, Evaluation and Examination Services.

Garret, H. E. (1966). Statistics in psychology and education (6th ed.). New York: David McKay.

Glass, G. V. (1978). Reply to Mansfield and Bussey. Educational Researcher, 7(1), 3.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. Review of Research in Education, 5, 351-179.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Researcher, 5(10), 3-8.

Glass, G. V., McGaw, B. & Smith, M. L. (1981). Meta-analysis in social research. Beverly Hills: Sage Publications.

Guilford, J. P. & Fruchter, B. (1973). Fundamental statistics in psychology and education (5th ed.). New York: McGraw-Hill.

Green, R. L. (1969). The black quest for higher education: An admission dilemma. The Personnel and Guidance Journal, 47, 905-911.

Gulliksen, H. & Wilks, S. S. (1950). Regression tests for several samples. Psychometrika, 15, 91-114.

Harris, A. D. (1940). Factors affecting college grades: A review of the literature, 1930-1937. Psychological Bulletin, 37, 125-166.

Harris, A. D. (1931). The relation to college grades of some factors other than intelligence. Archives of Psychology (No. 131).

Hedges, L. V. (1982). Statistical methodology in meta-analysis (ERIC/TM 83). Princeton, NJ: ERIC Clearinghouse in Tests, Measurement, & Evaluation.

Hedges, L. V. (1982b). Fitting categorical models to effect sizes from a series of experiments. Journal of Educational Statistics, 7, 119-137.

Hedges, L. V. (1982c). Estimation of effect size from a series of independent experiments. Psychological Bulletin, 92, 490-499.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and relevant estimators. Journal of Educational Statistics, 6, 107-128.

Hedges, L. V. & Olkin, I. (1983). Regression models in research synthesis. The American Statisticism, 37, 137-140.

Hedges, L. V. & Olkin, I. (1980). Vote-counting methods in research synthesis. Psychological Bulletin, 88, 359-369.

Humphreys, L. G. (1973). Statistical definitions of test validity for minority groups. Journal of Applied Psychology, 54, 1-4.

Humphreys, L. G. (1972). Implications of group differences for test interpretation. Proceedings of the 1972 ETS Invitational Conference in Testing Problems (pp. 56-71). Princeton,NJ.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills: Sage Publication.

Jackson, G. B. (1980). Methods for integrating reviews. Review of Educational Research, 50, 438-460.

Jensen, A. R. (1980). Bias in mental testing. New York: Free Press.

Kamp, L. J. T. V. D., Langerak, W. F., & Grkuijtter, D. N. M. D. (Eds.). (1980). Psychometrics for educational debates. London: Jonn Willey.

Kendrick, S. A. & Thomas, C. L. (1970). Transition from school to college. Review of Educational Research, 40, 151-179.

Kerlinger, F. N. & Pedhazur, E. J. (1973). Multiple regression in behavioral research. New York: Holt, Rinehart, & Winston.

Kleinbaum, D. G., & Kupper, L. L. (1978). Applied regression Analysis and other multivartiate methods. North Scituate, MA: Duxbury Press.

Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. Journal of Educational Statistics, 8, 93-101.

Lavin, D. E. The prediction of academic performance. A theoretical analysis and review of research. New York: Russell Sage Foundations.

Light, R. J. & Smith P. V. (1971). Accumulating evidence: Procedures, for resolving contradictions among different research studies. Harvard Eudcational Review, 41, 429-471.

Linn, R. L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A. K. Wigdor & R. G. Wendell (Eds.), Ability testing: Uses, consequences and controversies (Part II, pp. 335-388). Washington, DC.

Linn, R. L. (1975). Test bias and the prediction of grades in law school. Journal of Legal Education, 27, 293-323.

Linn, R. L. (1973). Fair test use in selection. Review of Educational Research, 43, 139-161.

Mosteller, F. M. & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (ed), Handbook of social psychology Vol. 1. Theory and Method (pp. 289-334). Cambridge, MA: Addison-Wesley.

Pearlman, K., Schmidt, F. L., & Hunter, J. E. (1980). Validity generalization for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 65, 373-401.

Pedhazur, E. J. (1982). Multiple Regression in behavioral sciences (2nd ed.). New York: Holt, Rinehart, and Winston.

Reynolds, C. R. (1982). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin, Handbook of school psychology (pp. 178-208). New York: John Wiley.

Rosenthal, R. (1978). Combining results of independent studies. Psychological Bulletin, 85, 185-193.

Rosenthal, R. & Rubin, D. B. (1979). Comparing significance levels of independent studies. Psychological Bulletin, 86, 1165-1168.

Samuda, R. J. (1975). Psychological testing of American minorities: Issues and consequences. New York: Dodd, Mead.

Scheunneman, J. D. (1981). A new look at bias in aptitude tests. New Directions for Testing and Measurement, 12, 3-35.

Schmidt, F. L. & Hunter, J. E. (1981). Employment Testing: Old theories and new research findings. American Psychologist, 36, 1128-1137.

Schmidt, F. L. & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. Journal of Applied Psychology, 62, 529-540.

Schmidt, F. L. & Hunter, J. E. (1974). Racial ethnic bias in psychological tests: Divergent implications of two definitions of test bias. American Psychologist, 29, 1-8.

Schmidt, F. L., Hunter, J. E. & Urry, V. E. (1976). Statistical power in criterion-related validity studies. Journal of Applied Psychology, 61, 473-485.

Schrader, W. B. (1971). The predictive validity of College Board Admissions Tests. In W. H. Angoff (Ed.), The College Board Admissions Testing Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and achievement tests (pp. 117-145). New York: College Entrance Examination Board.

Shuey, M. (1966). The testing of negro intelligence. New York: Social Science Press.

Siegel, S. (1956). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill Book.

Slavin, R. E. (1983, April). Meta-analysis in education: Meta-science or meta-nonsense? Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Slack, W. V. & Porter, B. (1980). The Scholastic Aptitude Test: A critical appraisal. Harvard Educational Review, 50, 154-175.

Smith, M. L., Glasss, G. V. (1977). Meta-analysis of psychotherapy outcome studies. American Psychologist, 32, 752-760.

Snecodor, G. W. & Cochran, W. G. (1980). Statistical methods (7th ed.). Iowa: Iowa University Press.

Stanely, J. C. (1971). Predicting college success of the educationally disadvantaged. Science, 171, 640-647.

Wagner, M. E. (1934). A survey of the literature on college performance prediction. University of Buffalo Studies, 9, 194-209.

Williams, R. L. Moderator variables as bias in testing black children. Journal of Afro-American Issues, 3(11), 77-90.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill Book.

Zar, J. H. (1974). Biostatistical Analysis. Englewood Cliffs, NJ: Prentice-Hall.

Zedeck, S. (1971). Problems with the use of "moderator" variables. Psychological Bulletin, 76, 295-310.

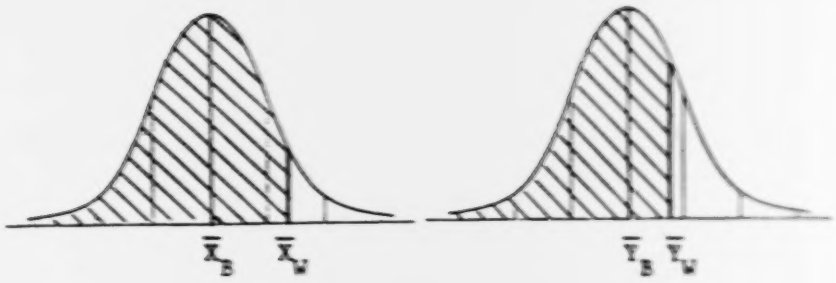

Figure 1 Blacks Mean Scores on the Predictor (a) and on the Criterion (b) Relative to Whites Mean Scores.